such as MEDLINE, synonymy causes reduced recall (that is, inability to find records that are truly related to the gene being sought) and polysemy causes reduced precision (that is, retrieving records that are not related to the gene, due to word sense ambiguities such as insulin being a gene, a protein, a hormone and a therapeutic agent). A more fundamental limitation is that the full text of the published biomedical literature contains far more information than is contained in the corresponding titles and abstracts of articles.

To overcome the problems of unpredictable word usage of authors, libraries and information scientists developed the notion of controlled vocabularies. Terms selected from 'preferred term lists' by professionally trained indexers (who, in the case of MEDLINE, are taught to read and apply keywords to the full text of the article *before* reading the abstract) have the potential to improve both recall and precision relative to searching of author-supplied text words. But even trained biomedical indexers reading the same article choose the same preferred keywords only 40–60% of the time[3], so a central issue in language-based systems is making them sufficiently robust and tolerant of the variety of ways of representing the biological ideas in the literature.

Associating experimental microarray results with the published literature is an example of a 'data mining' tool that uses explicit linkages between two independently constructed sources of information that contain conceptually related records. Successful data mining depends critically upon reliable sets of what computer scientists call "foreign key references"—that is, the existence of the same unique words or record identifiers in records of each of the sources to be linked. The PubGene application uses HUGO gene symbols associated with specific loci on microarrays as the common currency for linking to the literature, and displays characterizations of genes using the MeSH keywords from the literature written about those genes. At the University of California, San Diego, we have developed a similar application (http://www.array.ucsd.edu) for interpreting gene clusters that uses GenBank accession numbers as the common currency for linking to the literature and extends the notion of characterizing groups of genes through literature-derived keywords by placing those keywords in concept hierarchies.

Interpretive tools that are currently available for analyzing microarray results provide only a partial view of the relevant literature, as if gazing through a picket fence. As Jenssen *et al.* have pointed out, PubGene gene pairs identified by co-occurences within titles and abstracts accounted for only 45% of gene pairs described in the text of the records of Online Mendelian Inheritance in Man, and 51% of the pairs of proteins described in the Database of Interacting Proteins. Although this is far above a chance level of performance, it means that such approaches are useful primarily as tools of intellectual exploration and browsing, and not as comprehensive or definitive tools for global characterization of expression results. Additional limitations include the inescapable fact that expressed sequence tags and genes without associated publications do not participate in the analysis, and the more subtle bias that well-known, better-characterized genes are over-represented in the literature relative to newly discovered genes.

The usefulness of automated linkages to the literature in assisting in the interpretation of array data will improve as the literature expands and becomes increasingly available as electronic full text, and as computational tools for processing language become more powerful and robust. In the meantime, the view through a picket fence is clearly better than no view at all. □

1. Jenssen T.-K., Laegrid, A., Komoroski, J. & Hovig, E. *Nature Genet.* **28**, 21–28 (2001).
2. Johnson, S. Preface to *Dictionary of the English Language* (London, 1775).
3. Funk, M.E., Reid, C.A. & McGoogar, L.S. *Bull. Med. Library Assoc.* **71**, 176–183 (1983).
4. Masys D.R. *et al. Bioinformatics* **7**, 319–326 (2001).

Packaging paternal chromosomes with protamine

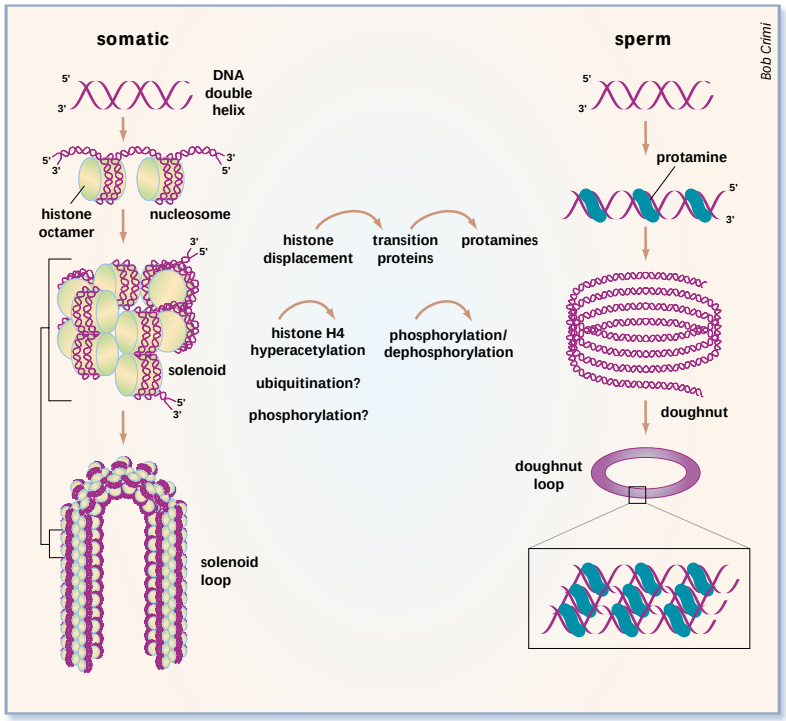
Robert E. Braun

Department of Genetics, University of Washington, Seattle, Washington 98195-7360, USA. e-mail: braun@u.washington.edu

The chromosomes of sperm cells are tightly packaged into a complex of DNA and protamines. Converting the chromatin from a nucleohistone to a nucleoprotamine structure may serve both biophysical and developmental functions. Several recent genetic studies have shown unexpected findings of the dosage requirements for the genes involved in sperm chromatin remodeling.

As spermatids undergo the terminal stages of spermatogenesis, compacting the DNA requires the replacement of the histones with a class of arginine- and cysteine-rich proteins called protamines. Although the reason for replacement of the histones with protamines is unknown, one possibility is the generation of a more hydrodynamic sperm head that speeds the transit through the female reproductive tract and across the zona pellucida surrounding the egg. It may also be that the nucleoprotamine structure protects the genetic material in the sperm head from physical and chemical damage. Alternatively, packing of sperm chromatin may serve to reprogram the paternal genome so that the appropriate genes from the father's chromosomes are expressed in the early embryo. Whereas most mammals contain a single protamine, mice and men have two. A study by Chunghee Cho, William Willis and colleagues[1] (see page 82) indicates that each is vital to paternal procreation.

Using gene targeting in mouse embryonic stem cells to investigate the function of the mouse protamine genes (*Prm1* and *Prm2*), Cho *et al.* show that mutation of either haploid-expressed gene leads to defective sperm. Unexpectedly, removal



Packaging chromatin. A model of chromatin packaging in somatic cells (left) and mammalian sperm (right). In somatic cells, the DNA is wound twice around histone octamers to form nucleosomes, which are then coiled into solenoids. The solenoids are attached at intervals to the nuclear matrix at their bases and form DNA loop domains. In the sperm nucleus, protamines replace the histones and the protamine-DNA complex is coiled into a doughnut shape. Inset shows the tight compacting of protamine-DNA strands. Displacement of the histones is facilitated by post-translational modifications of the proteins, in the form of histone H4 acetylation, ubiquitination and phosphorylation. Phosphorylation and dephosphorylation of the transition proteins facilitate their displacement before protamines bind.

of a single copy of either *Prm1* or *Prm2* is detrimental to postmeiotic spermatids, demonstrating haploinsufficiency. Both types of meiotic products—sperm carrying the mutant protamine allele and sperm carrying the wild-type allele—are nonfunctional. Presumably the wild-type sperm are affected as a result of the syncytial nature of spermatogenesis[2,3]. Incomplete cytokinesis during mitosis and meiosis generates clusters of haploid spermatids that remain connected through cytoplasmic bridges. The intercellular bridges connecting spermatids effectively equilibrate the 1× dosage of gene product produced by the wild-type spermatid, and the zero dosage produced by the mutant spermatid, to a 0.5× dosage distributed among all the cells. This reduction in protamine causes defects in DNA compaction.

Cho *et al.* are to be commended for their demonstration that protamine genes behave in a haploinsufficient manner. Because most ES cells are XY, and transmission of mutant alleles is accomplished by breeding chimeric males, the authors had to investigate the haploinsufficiency in chimeric males. Analysis of the mice required distinguishing between a mixed population of sperm containing three genotypes; sperm derived from the recipient C57BL/6N blastocysts, and sperm derived from the 129 Sv/J ES cells carrying either the wild-type or mutant protamine allele. The data show convincingly that wild-type and mutant 129 Sv/J sperm are made in equal numbers and that both are nonfunctional[1].

Prelude to protamines

In mammals, the protamines do not directly replace the histones. Instead, another group of basic proteins, the transition proteins, act as intermediaries in the histone-to-protamine transition. Mice have two major transition proteins, encoded by *Tnp1* and *Tnp2* (ref. 4). Experiments in which *Tnp1* is deleted indicate that the transition proteins have largely redundant functions[5]. Neither *Tnp1* or *Tnp2* alone are haploinsufficient, in fact, mutants homozygous for either gene are fertile, although their litter sizes are smaller, and sperm have relatively minor head abnormalities (M. Meistrich, pers. comm.). However, reduction of the total *Tnp* dosage by 75% in either *Tnp1*- or *Tnp2*- null mice lacking one copy of the other *Tnp*, or 100% elimination of both transition proteins in double-null mutants, results in more severe abnormalities in nuclear condensation and sterility. The similar phenotypes observed in all combinations of equivalent *Tnp* dosage indicate a common function for the two transition proteins.

Do the two mouse protamines also have redundant functions? The presence of a single protamine in most mammals would suggest that they do. If so, the haploinsufficiency is even more remarkable as it suggests that chromatin remodeling is compromised by as little as a 25% reduction in protamine levels (three functioning alleles in a four-allele system). However, several observations indicate that the protamines have evolved separate functions. First, the stoichiometry of the two protamines in humans and mice are different and, moreover, Cho *et al.* show that in $Prm2^{+/-}$ chimeras, there is also less Prm1 protein present in sperm, indicating that Prm1 deposition requires normal amounts of Prm2. In addition, in transgenic mice that overexpress Prm1 at its normal time during spermatid differentiation, the ratio of protamine 1 to protamine 2 in mature sperm is similar to that in wild-type sperm[6,7]. Somehow the cells can distinguish between the two pools of protamines and assemble a nucleoprotamine structure with wild-type stoichiometry. One wonders whether an extra gene dose of *Prm1* can substitute for the missing dose of *Prm2* and vice versa.

Nuclear condensation

Displacement of histones by transition proteins and protamines is accompanied by several post-translational modifications (see figure). Biochemical studies in mammals, fish and birds indicate that histone acetylation (especially histone H4 acetylation[8]), ubiquination and phosphorylation all facilitate the displacement of histones. Chromatin remodeling may also require chaperones that actively displace the post-translationally modified histones. Phosphorylation of the transition proteins and the protamines, presumably important for neutralizing these highly basic proteins, may also be important for chromatin compaction[9]. It was recently shown that targeted mutations in the HR6B ubiquitin-conjugating enzyme, *Ube2b*, and *Camk4* disrupt the terminal stages of spermatid differentiation and cause sterility. *Camk4* encodes the Ca^{2+}/calmodulin-dependent protein kinase IV (ref. 10), and phosphorylates Prm2 (ref. 11). Both genes are expressed in other tissues, yet have phenotypes that are restricted to the testis. The temporal and spatial pat-

tern of spermatid differentiation, the increasing power and sophistication of mouse genetics and the opportunity to perform biochemical phenotyping make spermatogenesis a highly attractive system for the study of chromatin repackaging.

Repack, reprogram

Developmental reprogramming of the parental genomes occurs during egg and sperm formation. However, the direct relationship between chromatin repackaging in sperm and developmental reprogramming is unknown. Fluorescence *in situ* hybridization indicates that chromosomes are not packaged haphazardly into sperm[12], and in humans, where about 15% of the DNA remains packaged in nucleohistones, it seems that specific sequences remain bound by histones[13]. Are the transition proteins and protamines directly involved in developmental reprogramming? Despite the sterility observed in the mouse *Tnp* and *Prm* mutants, many of the sperm look surprisingly normal. Have the nuclei present in these sperm undergone proper developmental reprogramming? A precedent for this class of mutants exists in flies and worms where certain paternal effect mutants generate sperm capable of fertilizing eggs, but defective in post-fertilization processes[14]. One of these paternal-effect mutants in *Drosophila*, referred to as *pal* (ref. 15) causes specific loss of the paternal chromosomes during the early embryonic cleavages following fertilization. Interestingly, *pal* encodes a small basic protein expressed late in spermatogenesis (B. Wakimoto, pers. comm.). The ability to fertilize mammalian eggs by intracytoplasmic sperm injection allows a direct test of the developmental potency of the *Tnp* and *Prm* mutant sperm nuclei. □

1. Cho, C. *et al. Nature Genet.* **28**, 82–86 (2001).
2. Braun, R.E., Behringer, R.R., Peschon, J.J., Brinster, R.L. & Palmiter, R.D. *Nature* **337**, 373–376 (1989).
3. Caldwell, K.A. & Handel, M.A. *Proc. Natl. Acad. Sci. USA* **88**, 2407–2411 (1991).
4. Meistrich, M.L. in *Histones and Other Basic Nuclear Proteins* (eds. Hnilica, G., Stein, G. & Stein, J.) 165–182 (CRC Press, Orlando, 1989).
5. Yu, Y.E. *et al. Proc. Natl. Acad. Sci. USA* **97**, 4683–4688 (2000).
6. Peschon, J.J., Behringer, R.R., Brinster, R.L. & Palmiter, R.D. *Proc. Natl. Acad. Sci. USA* **84**, 5316–5319 (1987).
7. Peschon, J.J. Thesis, Univ. Washington (1988).
8. Meistrich, M.L., Trostle Weige, P.K., Lin, R., Bhatnagar, Y.M. & Allis, C.D. *Mol. Reprod. Dev.* **31**, 170–181 (1992).
9. Green, G.R., Balhorn, R., Poccia, D.L. & Hecht, N.B. *Mol. Reprod. Dev.* **37**, 255–263 (1994).
10. Wu, J.Y. *et al. Nature Genet.* **25**, 448–452 (2000).
11. Roest, H.P. *et al. Cell* **86**, 799–810 (1996).
12. Ward, W.S. & Zalensky, A.O. *Crit. Rev. Euk. Gene Expr.* **6**, 139–147 (1996).
13. Gatewood, J.M., Cook, G.R., Balhorn, R., Bradbury, E.M. & Schmid, C.W. *Science* **236**, 962–964 (1987).
14. Fitch, K.R., Yasuda, G.K., Owens, K.N. & Wakimoto, B.T. *Curr. Top. Dev. Biol.* **38**, 1–34 (1998).
15. Baker, B.S. *Genetics* **80**, 267–296 (1975).